THE GENETIC CODE

F.H.C. Crick

It gives me very great pleasure to be here to give The Charles West Lecture, but it causes me a certain degree of embarrassment because I am, in origin, a physicist and this is strictly a medical lecture and a medical occasion. My excuse is that the topic I have chosen - the genetic code - is one which is of very great importance to all biological organisms, ourselves included. If you were to press me and ask exactly how our knowledge of the genetic code will be important for medicine I could not easily, at this stage, tell you precisely, but I can make the general forecast that it will certainly have repercussions in medicine within our lifetime.

The problem of the genetic code is basically speaking the problem of protein synthesis. Proteins are important for two main reasons. The major reason is that all the catalysts that we find in living cells - all the enzymes - are made of protein. In addition, in animals, but less so in plants, proteins also fulfil a structural role. For a simple organism, such as a bacterial cell, it is almost true to say that once you have specified the correct proteins you have done the job of specifying the whole organism. Of course it is true that you have got to start with a cell. That is, you must have pre-existing components of various kinds, but the most important thing is to get all the proteins right. We do not know exactly how many different proteins there are in, for example, Escherichia coli, but a reasonable guess would be perhaps two thousand.

Let us consider briefly the elementary chemistry of proteins. Proteins are an enormous family of molecules, but they are all built on the same general plan. Each one consists of a polypeptide chain, having a regular repeating sequence of atoms, with side chains attached at regular intervals. You could hardly have a simpler basic structure. However these side chains are not all the same. Each protein has a characteristic length of chain, usually several hundred residues long, but there are only twenty different kinds of side chains commonly found in proteins. A protein is in fact made by joining together amino-acids with the elimination of water. The twenty kinds of amino-acid are joined together - head to tail - to give long polypeptide chains. Each particular protein - your own haemoglobin, for example - has a particular sequence of these amino-acid residues along its backbone.

Proteins, therefore, from this point of view are remarkably simple. How does it come about then that they are so important? They are important because, as I have said, they act as highly specific catalysts and they do this because they can form intricate and subtle three-dimensional structures. I have spoken as if the chain was fully extended, but this is not so. In general, the chain folds up into helices which then fold on themselves to make ~~complicated~~ complicated cavities lined with active chemical groups. The protein myoglobin, for example, has a complicated three-dimensional structure which holds the haem group, whose function is to carry the oxygen molecule. It is an interesting fact that this enormous structure is needed just to hold a small molecule of oxygen. The reason for this is quite simple - proteins are the main things the cell knows how to make. Everything else has to be made by proteins. It is much more difficult to synthesize the relatively small haem (because it requires several enzymatic

steps to do this, and for each one a special protein is required) than to make the protein part of myoglobin.

We have now looked at proteins from two points of view. From the point of view of their chemical structure we see that although they are large molecules, they are simple. They are made from a set of twenty standard types of unit, - twenty amino-acids - joined together in a regular way, head-to-tail. On the other hand from the point of view of their physical shape, they are complicated and intricate molecules. We believe that each protein obtains its own particular shape by folding itself up, and there is now suggestive evidence for this idea. One of the great secrets of biochemistry is this ability of make complicated catalysts by a relatively simple method.

The key problem, therefore, in obtaining a particular protein is to join up the amino-acids in the correct order to give a polypeptide chain which will then fold itself up into the ~~correct~~ required three-dimensional structure. We now know that this is the most important function of the genetic material. Its main function - certainly in simple organisms - is to determine the structure of proteins. The old slogan of Beadle 'one gene, one enzyme' expresses this particular point of view.

It is now reasonably sure that genes are made of nucleic acid. We only know this quite certainly, I think, for a few viruses, and for one or two bacteria. We surmise that it is true for higher organisms. Nucleic acid is of two sorts - DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). The main genetic material is in fact DNA. Once again, when we look at its chemical structure we see that it, too, is a polymer. It consists of a very long, regular backbone which goes phosphate-sugar-phosphate-sugar-phosphate-sugar and so on indefinitely. To each sugar, at the same point, is attached a little side

group called a base, of which there are four sorts: adenine, cytosine, guanine, and thymine. RNA is the genetic material of certain viruses, such as poliomyelitis, and also has other functions important in the cell for protein synthesis. It is very similar to DNA. The main difference is that the sugar is ribose rather than deoxyribose, and that it contains uracil instead of thymine. The essential point is that for both DNA and RNA the backbone is regular and is made by joining groups together head-to-tail, though for nucleic acid, as opposed to protein, there are only four different types of group commonly found in Nature.

How long, typically, is a molecule of nucleic acid? This is not an easy question to answer. It suffices to say that for genetic material it is usually at least several thousand residues long and may be very much longer than this. Sometimes the molecules are in the form of a double helix, in which two chains fit together in an interesting way. This is usual for DNA. Sometimes, especially for small viruses, the molecule consists of a single chain. The little RNA viruses - like poliomyelitis - are usually single-stranded RNA, and are often about five or six thousand residues long. Viruses with double-stranded RNA also occur, and also ones with single-stranded DNA. Some viruses are bigger than this. For example bacteriophage T4 has double-stranded DNA as its genetic material. The DNA is all in one piece and is about 200,000 residues long. It probably has of the order of 100 separate genes.

We can go to a scale larger and ask about a bacterial cell, such as E. coli. It is one of the more recent and surprising discoveries that the DNA of such a cell - the genetic material - appears to be all in one long piece having about 3×10^6 base pairs. The actual length of this DNA is about a millimetre.

This should be compared with the size of the bacterial cell, which is about 1 or 2μ. There is quite a lot to pack in, since a millimetre of the genetic material has to go inside a cell which is only about a thousandth of a millimetre in length. For man the total length of the DNA is even greater. If you were to take one of your own cells and put all the DNA molecules in it end-to-end they would reach a distance of about two metres.

What, then, is a gene? A gene, on this picture, is a certain length of a molecule of nucleic acid. We believe that the genetic message is expressed by the sequence of the four different bases. In the morse code you have a dot, a dash and a space. In nucleic acid you have four symbols and the exact order of these symbols we believe carries the genetic message. There must be signals for saying what is the beginning and the end of a gene but we believe that they, too, are written in this language of the four letters. So a gene is a stretch of nucleic acid, typically some thousand or so bases long.

The function of a gene is to instruct the cell how to make a particular polypeptide chain of a particular protein. The Sequence Hypothesis says that it does this because the sequence of bases in the nucleic acid corresponds in some simple way to the sequence of the amino-acid of the protein. Put another way, we have to ask the question - how is it that a sequence of four things can control the sequence of twenty quite different things? This is like asking, for the morse code, how can the sequence of dots and dashes be translated into an alphabet of twenty-six letters? The expression "The Genetic Code" means not the message itself but the way that you translate from one language to the other.

You might think that this would be a straightforward

problem to solve. For certain favourable proteins it is possible to determine their amino-acid sequence. For example, the amino-acid sequence of ribonuclease, the enzyme from the pancreas which helps to digest the nucleic acid in the intestines, has been determined. Another example is the protein of Tobacco Mosaic Virus, which makes up the shell of the virus. Surely, you might think, it is not too difficult to determine the amino-acid sequence of a protein and in addition determine the base sequence of the gene which controls it, and then compare the two. The relationship between the two sequences would no doubt be transparent.

Unfortunately it turns out one cannot easily do this at this moment. In order to determine sequence you must have a fairly large supply of whatever it is you are looking at - a particular protein, for example. It must be pure, or at least reasonably pure. There must be methods for breaking it down in a controlled manner so that you can do your problem in bits. For nucleic acid that is not easy. We do not have a simple method yet of purifying a single gene. Until recently the best we could do was to work on a small virus, which has between five and ten genes, say, and which has about six thousand bases in its RNA. That is quite different from working on a sequence with only a few hundred residues. Moreover, nucleic acid chemistry is rather more difficult than protein chemistry. It also turns out that it is much more difficult to determine a sequence of four things than a sequence of twenty things. In a sequence of twenty things one or two of them are probably rather rare and this makes it easier to determine the sequence. In a sequence of four things one does not have that help. Consequently the straightforward, direct approach, outlined above, cannot be used.

We have another technique, however, which we can use. This is the technique of genetic mapping. From classical genetics we know that in higher organisms genes are arranged in a linear order along a chromosome and this order can be determined by studying their rate of recombination. The further two genes are apart, the more often they recombine, and so by measuring this frequency of recombination one can measure a genetic map distance. One of the things which one would infer from what I have said of the genetic material is that you would also expect to have genetic recombination between two markers within a single gene, and would therefore be able to make a genetic map of the sites within a gene. Naturally these sites would be very close together, and consequently recombination between them would be very rare. To pick it up one has to use very large populations, and in fact it is only technically possible in a worthwhile way using micro-organisms, where one can handle very large populations and where it is possible to use selective techniques to pick out the rare recombinants one is trying to find.

By this method it has indeed proved possible to make a map of the order of sites within a single gene. This has been done by Benzer using the bacteriophage T4. He has found as many as several hundred sites within the pair of genes on which he works. (Whereas in the past people used to concentrate on one organism and perhaps, if they got a little specialised, on one chromosome, now people concentrate on one gene and I myself only work on the left-hand end of one particular gene). It was also shown by Benzer (1959 and 1961) that all these sites were arranged in a linear order.

It is convenient at this point to introduce the word

"codon" - which means the group of bases on the nucleic acid which codes for one particular amino-acid. As we shall see, it is likely that each codon consists of three bases. If we could write down the codons corresponding to each of the twenty amino-acids such a list would define the genetic code, and that would be the end of my lecture. Unfortunately although molecular biology has advanced very rapidly it has nor gone quite as fast as that. We know something about the genetic code, but we do not yet know it in full detail. So we have to start off by asking a certain number of general questions of which the first very pertinently is - what is the size of the codon? Obviously it cannot be just a pair of bases. We have four alternatives for each site, and this would give sixteen combinations, which is not enough to code for twenty different things. In addition these must be a few symbols for spaces, such as stop and start. So the smallest likely number, assuming that it is an integer which is always the same, is three bases per codon. This gives us sixty-four different possible triplets and this, as you will see, we now think is the correct answer.

This leads us to a second question. Will there be more than one codon standing for each amino-acid? You could imagine that out of the sixty-four only twenty stood for amino-acids and forty-four of the triplets stood for something else. We do not believe that this is so. The present evidence suggests that, in general, more than one triplet codes for a particular amino-acid. The code, to use the technical term, is degenerate.

Now the four bases, with minor reservations, and the twenty amino-acids, also with minor reservations, are the same right throughout Nature. They are the same for animals, for plants and for micro-organisms. For this reason we can attack the problem of the genetic code (to start off with at any rate)

by studying which organism we please. Not unnaturally, we choose the ones which we can handle most easily. It turns out that the ones most used are, on the one hand, micro-organisms, because one can grow them so rapidly, and their nutritional requirements are easy; but also, I am glad to say, human beings. This is because there are so many doctors looking at so many human beings.

Let us assume for the moment that it is a triplet of bases which makes a codon. There are two main possibilities open to us. One is called an overlapping code, and the other a non-overlapping code. The distinction is quite simple. In the non-overlapping code the first amino-acid is coded by the first three bases, the next by the next three, and so on. In an overlapping code, on the other hand, the first amino-acid is coded by base one, two and three but the second one by numbers two, three and four, and third by numbers three, four and five, etc. We are now reasonably certain that the non-overlapping code is right. The reason is as follows: a genetic mutation will in general change one base into another base. It will alter the message at one particular point. With a non-overlapping code, this would mean that you would merely alter one amino-acid. However, the typical change in an <u>overlapping</u> code should be to several adjacent amino-acids. This is never found. Our evidence comes mainly from the numerous abnormal human haemoglobins and from the protein of Tobacco Mosaic Virus, (Wittmann and Wittmann-Liebold, 1963; Tsugita, 1962).

We ask next: is the order of the codons along the nucleic acid the same as the order of the corresponding amino acids in the protein? In other words, are the two orders co-linear? Experimental evidence that this is so has been produced recently by two groups of workers.

Yanofsky and his co-workers (1964) have studied one of the proteins of the enzyme tryptophan synthetase of Escherichia coli. They have determined part of the amino-acid sequence, by the methods of protein chemistry. In addition they have picked up a number of mutants, each of which, as I have already said, ~~have~~ has a change of one amino-acid. For example, one of them has a cystine instead of a tyrosine at a certain point, while another has an isoleucine irstead of a threonine, and so on. By genetic methods (which have nothing to do with protein chemistry) they have constructed a genetic map of these mutants, and put them in a certain order by means of genetic crosses. The order of the mutants on the genetic map is precisely the order of the corresponding amino-acid changes in the protein. Moreover, where the distance between two mutants is large, it is also large on the genetic map; and conversely, when they are very close together in the protein, they are also close together on the genetic map.

Secondly, some of my colleagues at Cambridge (Sarabhai, Stretton, Brenner and Bolle, 1964) have come to a similar conclusion, again using a micro-organism, but this time the bacteriophage T4. They have studied the protein of the head of this phage. They have picked up mutants of a particular type known as "amber" mutants, and have proved that such mutants terminate the polypeptide chain. You might wonder how can you work with a phage if its head is incomplete. The answer is that there are certain strains of E. coli (the host cell of the phage) which suppress such mutants, so that one can use these strains to grow a stock of the phage and then test how the mutant behaves by growing it on the usual strain. By relatively simple methods, using radioactive labelling, they have been able to estimate which of ~~the~~ any two different polypeptide fragments is the longer. For example, there is one particular peptide, containing cysteine,

which can be obtained by cutting up the protein with proto-lytic enzymes. (There is only one cysteine residue in the protein). One can ask, for each one of the mutants, whether it has got that cysteine peptide or not. The ones that have it will have longer chains than the ones that lack it. This general method can be applied to all the other peptides into which the protein can be split. The polypeptide chain that gives the most fragments is clearly the longest and therefore by looking at the evidence in this way, without determining the sequence precisely, they can determine the order of the mutants purely by the methods of protein chemistry. They have also obtained the order of the mutants by genetic methods, and, once again, the two orders are the same.

The next question is the following: since we know that we are reading along in groups of a certain number, which do not overlap, how do we know how to cut the message up into the right groups? For example, if you started reading in the middle you might begin out of phase- you might be one ahead or one behind, and then you would read the whole of the message wrongly. It turns out that we now have good evidence that it is done in the simplest way you could think of. The reading begins at some fixed point - you might say the beginning - and just goes on reading from there in groups of three. I will not go into the evidence in detail except to say that we have studied a deletion that joins two genes together and from this we can make a strong case that the reading does start from a particular point, (Crick, Barnett, Brenner and Watts-Tobin, 1961).

This has enabled us to tackle the question: is it really a group of three that makes up a codon? The basic idea is the following. We are able to pick up mutants which we believe (from the way they behave in various contexts) are not merely the change of one base into another, but are either the addition

or a deletion of a base or bases. What happens when you have a genetic message and you put in an extra base? The reading starts from the beginning until it comes to that point and from there onwards the whole of the message is read incorrectly, because it is being read out of phase. In fact we find that these mutants are completely inactive - this is one of our bits of evidence that they are what we say they are. You can pick up a number of such mutants and can put them together, by genetic methods, into the same gene. For example you can put together two of them. Such a gene would be read correctly until it reached the first addition, and then it would be out of phase. When the reading came to the second addition it would read out of phase again, and so the whole of the rest of the message would be read incorrectly. Now it so happens that the left-hand end of this gene is not terribly important for its function. We can actually delete it and the gene will work after a fashion. In this region we have constructed, by genetic methods, a <u>triple</u> mutant, using three mutants all of the same type, and we have found that the gene will nevertheless function fairly normally.

This result is really very striking. Each of the three different faults, used singly, will knock out the gene. You can put them together in pairs in any combination you like, but then the gene is still quite inactive. Put all three in the same gene and the function comes back. We have been able to do this with a number of distinct combinations of three mutants (Crick et al., 1961).

The conclusion we draw from this is that there are indeed three bases in a codon. I would emphasize that this is genetic evidence and it has to be confirmed by biochemistry before we can be certain of it.

Unfortunately I cannot go any further without saying just a few words about protein synthesis. We have been talking up till now as if protein synthesis were a black box operation, but sooner or later it is obvious we have to come down to biochemistry and ask what exactly is going on inside the cell. It will be impossible for me in this lecture to make you understand in any detail what our knowledge of protein synthesis amounts to, because it is now quite considerable, (though still a long way short of what we would like) but I will attempt to describe it in outline. For details see the review by Watson (1964).

The genetic material in most organisms is DNA. Proteins are made on objects called ribosomes, which are small particles (about the size of a polio virus) consisting of two unequal parts. We believe that they are the reading mechanism. They do not actually work on the DNA itself. A RNA copy is made of one chain of the DNA. This copy is called messenger RNA (m-RNA). The reading heads (the ribosomes) go onto the m-RNA and travel along it, one after another, giving simultaneous readings (one behind the other) of this working copy of the genetic message. As they go along they construct the protein and when they come to the end (assuming it is a simple message and only codes for one protein) they come off, and release the protein, and then go round again, not necessarily onto the same piece of messenger RNA. The string of ribosomes attached to a piece of m-RNA is known as a polyribosome or polysome for short. They can be seen in the electron microscope.

How do the amino-acids, under the dictates of the m-RNA, get into the right places? There is elaborate biochemical machinery to do this. Each amino-acid is carried in by one of a special species of molecule, known as soluble RNA (also called transfer RNA).

There is a special enzyme for each amino-acid which joins it specifically to its own soluble RNA and not to any of the others. There are twenty different sorts of enzymes - one for each amino-acid. However there are more than twenty kinds of soluble RNA, and therefore there are several different sorts for each amino-acid. We think there is probably one for each codon.

When the amino-acid has been added to the polypeptide chain the soluble RNA is then released, recharged with a new copy of the amino-acid and used again. You can see that the system is quite elaborate.

The important point I want to make is that a single-stranded RNA can carry the instructions to make a protein. This has been shown very convincingly in the case of RNA from a small virus, by adding it to a cell-free system (Nathans et al., 1962). This cell-free system is obtained by taking cells, breaking them open, partially purifying the bits and then putting them together again. To such a system you can add synthetic nucleic acids and they, too, will act as messengers. The recent work on the genetic code has gone very rapidly because of this technique. The initial discovery was by Nirenberg and Matthaei (1961) who added polyuridylic acid to such a system, and found that it responded by making polyphenylalanine, showing that one of the codons for phenylalanine consists of three uracils. His group and Ochoa's group have done most of the experimental work using this method. They have also shown that polyadenylic acid will direct the synthesis of polylysine, implying that three adenines code for lysine. In the same way it is likely that three cytosines code for proline, though this is less certain.

It is not too difficult, as it turns out, to make synthetic messenger RNA of determined composition but with a random sequence, because there is a special enzyme which will join the

precursors together. It is thus possible to synthesize an RNA which has uracils and cytosines in an approximately random order. This is found to icnorporate phenyalanine (UUU), proline (CCC) and also leucine and serine. Thus some of the triplets containing U's mixed with C's code for leucine and serine. By these methods, without going into details, they have been able to obtain the composition (but not the order) of quite a number of triplets. In certain cases more than one triplet is claimed for an amino-acid. We believe as a result of this, and other evidence that it is very probable that more than one triplet codes for a particular amino-acid, (see the review by Crick, 1963a and the articles in CSHS, 1963).

Recently Leder and Nirenberg (1964) have found a way of determining codons not by using long messages but just by adding triplets (trinucleotides) to the system, by studying the binding of s-RNA to ribosomes in the presence of a particular trinucleotide. They have been able to show that the triplet GUU (guanine, uracil, uracil, in that order) codes for valine and that UGU and UUG do not. Incidentally this gives us direct biochemical evidence that the codon is probably a triplet. These results will have to be extended and confirmed, but it does look as if we have a chance now of determining the order of all the triplets making up the code, though the results will need to be supplemented by other methods, of which mutagenesis is probably the most important.

It is interesting that it is already possible to obtain genetic recombination <u>within a codon</u>. For example, Yanofsky (1963) has studied a particular protein, which has glutamic acid at one place in the protein, in one mutant, and has argenine (in the same place) in another mutant. He has crossed these

together genetically, and has obtained glycine at that particular place (which is the amino acid found in the wild type). Genetic recombination has produced an amino-acid which was not there before. This is not surprising if you think of it in terms of the bases of the nucleic acid. The triplet in one mutant has been altered (from the wild type) in one place and in the other mutant in a different place. The recombinant has produced the original triplet again. This sort of result will enable us to confirm the genetic code when we have a tentative version of it.

I have said that the four bases are universal (the same throughout Nature), and that the twenty amino-acids are also the same throughout Nature. One might guess from this that the code (the relationship between them) might also be the same, but this has to be proved. We now have preliminary evidence about this. We can make cell-free systems from different organisms and add synthetic messengers to them and see if the same polypeptide chains is made in each case. So far nobody has shown a significant difference between one organism and another (see, for example, Weinstein, 1963). Another technique is to use mixed systems. It is possible to synthesize haemoglobin in the test-tube, using a cell-free system from reticulocytes. In addition it can be done using the polysomes from the reticulocytes, and part of the soluble components - the s-RNA and the activating enzymes - from E. coli. In other words a mixed system from a mammal and a micro-organism will synthesize a mammalian haemoglobin (von Ehrenstein and Lipmann, 1961).

The most dramatic case of a mixed system was discovered recently by Abel and Trautner (1964), who succeeded in growing vaccinia virus inside Bacillus subtilis. They infected the bacterial cell with the DNA of the vaccinia virus, and the cell became stuffed full of vaccinia. It is indeed a very remarkable thing that you can take the genetic material of an animal virus

and use it to grow that virus inside a micro-organism. This would again suggest that the genetic code of these two organisms must be very similar.

In one's enthusiasm for the genetic code one must realise that, apart from such details as to how you start the polypeptide chain and how you finish it, there are very important problems which are not strictly touched on by the code. Genes have to be turned on and off - they do not act all the time. This is a problem which is not strictly part of the genetic code, although it lies very close to it. We do not yet understand the control mechanisms in bacteria. We do not even know them in the case of phage infection. In the case of phage T4 there are early proteins which are made in the first ten minutes. A whole lot of other proteins (the proteins which make up the virus) are made later on. What controls the switching from early to late proteins we do not yet know. Still less do we know about higher organisms, or even if the control systems used in micro-organisms also exist in man. It is very likely that there are in addition more elaborate and heirarchal controls which occur in higher organisms.

Leaving this aside and coming back to the code, what can we say about it in summary? On the present evidence (which I must emphasize is incomplete) it appears to be a non-overlapping triplet code, with several triplets for each amino-acid. The code may not be identical in all living things but it is likely to be fairly similar.

The final question I think you might ask me is, when are we likely to know it in detail? This is always the most difficult question to answer but I would be surprised if it had not been determined, at least for one organism, within the next couple of years.